D110–D117 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 17 November 2012*
doi:10.1093/nar/gks1058

APPRIS: annotation of principal and alternative splice isoforms

Jose Manuel Rodriguez[1], Paolo Maietta[2], Iakes Ezkurdia[2], Alessandro Pietrelli[1], Jan-Jaap Wesselink[2], Gonzalo Lopez[2], Alfonso Valencia[1,2,*] and Michael L. Tress[2,*]

[1]Spanish National Bioinformatics Institute (INB) and [2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

Received August 14, 2012; Revised October 9, 2012; Accepted October 11, 2012

ABSTRACT

Here, we present APPRIS (http://appris.bioinfo.cnio.es), a database that houses annotations of human splice isoforms. APPRIS has been designed to provide value to manual annotations of the human genome by adding reliable protein structural and functional data and information from cross-species conservation. The visual representation of the annotations provided by APPRIS for each gene allows annotators and researchers alike to easily identify functional changes brought about by splicing events. In addition to collecting, integrating and analyzing reliable predictions of the effect of splicing events, APPRIS also selects a single reference sequence for each gene, here termed the principal isoform, based on the annotations of structure, function and conservation for each transcript. APPRIS identifies a principal isoform for 85% of the protein-coding genes in the GENCODE 7 release for ENSEMBL. Analysis of the APPRIS data shows that at least 70% of the alternative (non-principal) variants would lose important functional or structural information relative to the principal isoform.

INTRODUCTION

Genes can generate a wide range of mature RNA variants through the alternative splicing process (1,2). In fact, studies have revealed that virtually all multi-exon human genes (3,4) are capable of producing at least two RNA transcripts by alternative splicing. Alternative splicing events that occur within coding regions will produce alternative transcripts that have the potential to be translated into distinct gene products. Those alternative transcripts that are not picked up by the cellular surveillance machinery, such as nonsense-mediated decay (NMD; 5,6), non-stop decay (7) and no-go decay (8), may contribute to an increase in the complexity of the cell. Alternative gene products often have a surprising level of diversity (9) and can therefore have very different biological and cellular properties. Thus, the suggestion that the rearrangement of exons conducted by alternative splicing may enrich the repertoire of cellular functions (10).

Genome annotation projects are producing gene sets saturated with alternative splicing variants (11). If alternative splicing does have the potential to expand the cellular functional repertoire in eukaryotic species, it would seem to be important to assign roles to these splicing variants. However, the sheer quantity of genomic data generated by these projects (12,13) presents serious challenges for functional annotation. At present, almost all the experimental information related to alternative coding variants has been generated for RNA transcripts rather than protein isoforms. Despite the fact that there is only piecemeal experimental data for the cellular role of alternative isoforms, it is possible to predict the likely biological effects of alternative splicing.

APPRIS has been developed within the GENCODE consortium (14) to annotate alternative gene products with reliable, biologically relevant data. GENCODE provides high-accuracy manual annotations of protein-coding loci and alternative variants as part of the ENCODE project (12,15). The GENCODE annotations are gradually replacing and augmenting the Ensembl (16) automatic annotations. As part of the GENCODE annotation process, APPRIS flags isoforms with likely altered structure, function or localization, and exons that are evolving unusually. The information from APPRIS is fed back to the manual annotators and has lead to the annotation of new isoforms.

*To whom correspondence should be addressed. Tel: +34 91 732 80 59; Fax: +34 91 224 69 76; Email: valencia@cnio.es
Correspondence may also be addressed to Michael L. Tress. Tel: +34 91732 8000; fax +34 91 224 69 76; Email: mtress@cnio.es
Present addresses:
Alessandro Pietrelli, Institute for Biomedical Technologies ITB—CNR, Milan 20900, Italy.
Jan-Jaap Wesselink, BIOMOL Informatics, S.L., Campus Universidad Autónoma de Madrid, Madrid 28049, Spain.
Gonzalo Lopez, Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA.

APPRIS annotates variants with protein structural and functional information, signal peptides and *trans*-membrane helices, conservation of related species, the conservation of exonic structure and exon evolutionary rates. APPRIS currently annotates the GENCODE/Ensembl merge of the human genome.

The novel feature of APPRIS is that it selects a principal isoform for each gene based on the reliable annotations for protein structure, function and cross-species conservation. The principal isoform is the representative isoform of the gene, the isoform against which all other (alternative) isoforms should be compared. In APPRIS, the principal isoform is the isoform with the main cellular function, the isoform that is expressed in the majority tissues or in most stages of development or the isoform that is the most evolutionary conserved. The process of selecting principal isoforms is illustrated with biologically relevant examples in the 'APPRIS annotation' section.

It is particularly difficult to automate the selection of a single representative for a gene, and all large-scale genomic analyses and databases such as Ensembl (16) and SwissProt (17) get round this problem by simply selecting the longest isoform as the main variant. Although this is a safe choice, and is often correct, we have shown that it is not always the best strategy—only ~75% of the isoforms selected by this strategy are likely to be principal (18).

We performed an initial study on the feasibility of selecting principal isoforms using a number of prediction methods (18). The methods used to pinpoint principal functional isoforms were based on conservation and the characteristics of known proteins, principally structural and functional features. We determined a principal variant for 179 of the 215 human genes in the study, 83% of genes with multiple alternative variants. Where the principal variants selected in the study differed from the SwissProt display sequences, we found annotation evidence from cross-species alignments that supported our selection over the SwissProt display sequence.

Based on this initial study, we developed APPRIS. APPRIS is made up of eight separate annotation modules, each with a specific role. For example, firestar (19) predicts the presence of individual functionally important residues in splice variants, Matador3D predicts the effect of splicing events on 3D structure and INERTIA detects exons that are undergoing unusual evolution.

THE DATABASE

APPRIS was developed using version 3c of the GENCODE annotation (Ensembl 56), which was the initial Ensembl/GENCODE merge, and currently runs GENCODE version 7 (Ensembl 62). Between GENCODE 3c and 7, the annotation was cleaned of ~2000 genes (mostly automatic annotations removed by Ensembl), while more 10 000 new annotated coding variants were added. GENCODE release 7 recognizes 20 687 protein-coding genes and 84 408 distinct-coding transcripts. APPRIS is updated with each new stable GENCODE release and is currently being updated to GENCODE version 12.

The APPRIS system (Figure 1) is composed of eight separate annotation modules. These eight modules do not comprise an exhaustive list of all possible protein features. Instead, the methods used in APPRIS were chosen for their ability to select principal isoforms. Each method either detects the absence of highly conserved protein features (as highly conserved protein features are extremely unlikely to have arisen by chance, we can discard variants that lack these features) or calculates cross-species conservation (the more conserved an exon/transcript, the more likely that it represents the principal variant). None of the computational methods behind each module is previously unpublished. Instead, the methods have either been combined in novel ways or have been adjusted especially for APPRIS, and the output of all the methods has been tuned to keep false-positive predictions to a minimum, albeit at the expense of coverage.

The eight modules are as follows (see Supplementary Data for further details). Matador3D checks for the presence of structural homologs in the PDB (20) and tests the integrity of the 3D structure; firestar (19) makes highly reliable predictions of conserved functionally important amino acid residues; SPADE uses the program Pfamscan (21) to count conserved and compromised Pfam functional domains; INERTIA uses three alignment methods (22–24) to generate cross-species alignments, from which SLR (25) identifies exons with unusual evolutionary rates; CRASH makes conservative predictions of signal peptides using the SignalP and TargetP programs (26); THUMP generates conservative predictions of *trans*-membrane helices from three separate *trans*-membrane predictors (27–29); CExonic uses exonerate (30) to align mouse and human transcripts and looks for patterns of conservation in exonic structure and CORSAIR uses BLAST (31) to map vertebrate orthologs to each variant and counts the numbers of orthologs that align correctly and without gaps. All of these methods are available as web services.

Annotation and selection of principal isoforms

In addition to annotating alternative isoforms with biological data, APPRIS selects a principal isoform from among the isoforms annotated for each gene. The selection of a principal isoform by APPRIS is based on two principles. The first principle is that there is often one isoform that performs the main cellular function or that is expressed in the majority tissues or in most stages of development, and that the rest of the annotated isoforms are alternatively spliced isoforms that may perform distinct roles. Proteomics evidence suggests that this seems to hold true for many genes (32–37) although there are genes for which it is more difficult to define a principal isoform precisely because there are a number of isoforms that might be regarded as equally important for cellular function.

The second principle is that the principal isoform should have more evolutionary history. The principal

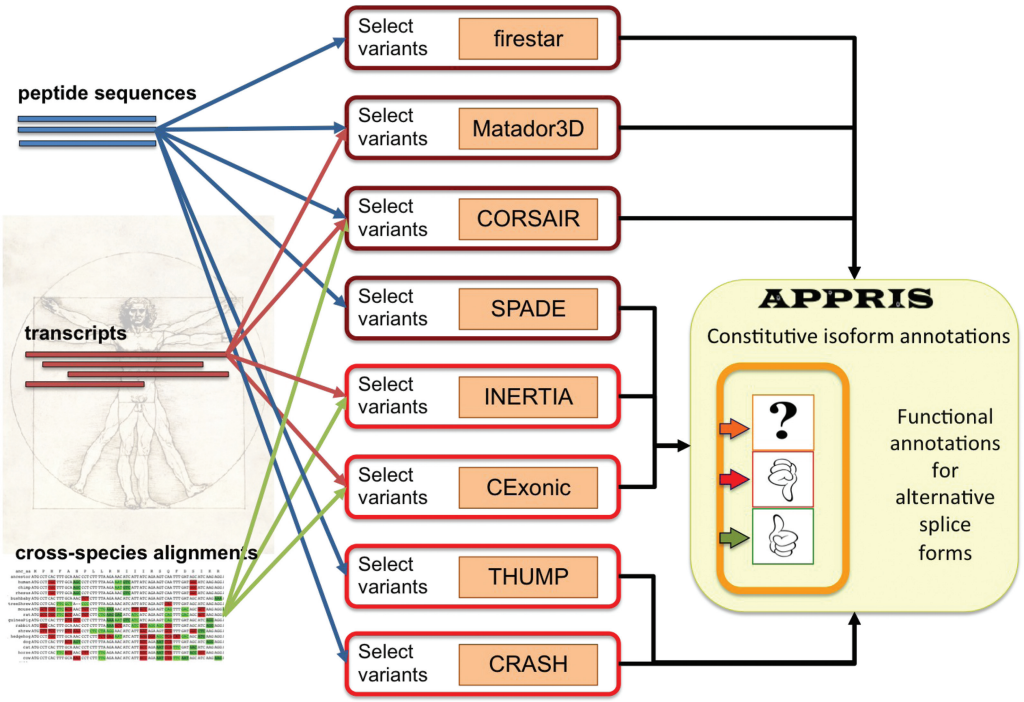


Figure 1. The APPRIS system. The inputs to the web services are the peptide sequences of the isoforms, the sequences of the transcripts and cross-species alignments of either transcripts or peptides. The output of all eight individual web services is used by APPRIS to annotate alternative variants with structural, functional and conservation information and to select a principal isoform for each gene.

isoform ought to be the variant that is most conserved across related species. It has been shown that alternative exons tend to have evolved more recently (38), so this is a reasonable assumption.

The methods that make up APPRIS detect unusual, missing or non-conserved features and will flag these transcripts as alternative. The selection of a principal isoform is based on a jury of the eight methods that make up the pipeline. The isoform selected as principal will be either the variant that has the most conserved protein features (since it is much more likely that alternative isoforms have lost rather than gained protein features such as 3D structure and function) or that has more evidence of cross-species conservation, or, most frequently, both. Four methods (SPADE, CORSAIR, Matador3D and firestar) make up the core of the jury system, with the other methods becoming more important in cases where these four methods are not able to make a decision.

It should be noted that GENCODE-coding transcripts are not all considered equally. First, transcripts with identical CDS (in other words those that undergo alternative splicing only in 3′- or 5′-UTR) are regarded as identical for the purposes of selecting a principal protein isoform in APPRIS. Second, transcripts that are annotated as NMD targets are annotated with protein features, but cannot be selected as the principal variant by APPRIS. The same is also true of all transcripts annotated as fragments.

For the few cases where the methods in APPRIS tag all the variants as alternative (in most cases these are genes with 'read-through' transcripts), the gene is brought to the attention of the GENCODE manual annotators.

A list of principal isoforms selected by APPRIS for each version of GENCODE/Ensembl is available. In the few cases where APPRIS is not able to determine the main isoform, the variant with the longest protein sequence is selected from among those isoforms not rejected by APPRIS.

APPRIS will be updated with the Ensembl/GENCODE database and will be extended to cover mouse gene models in the near future as the GENCODE annotators focus on mouse models, although in theory APPRIS could be extended to incorporate data from any well-annotated eukaryotic genome.

System architecture and user notes

The APPRIS web site allows the user to search genes and transcripts and displays six panels of annotations. The first panel shows all the GENCODE/Ensembl-coding variants and highlights the main functional variant. The second panel shows the APPRIS annotations in detail and includes information such as the number of functional residues detected by firestar, the Matador3D homologous structure score or the number of vertebrate species that align in CORSAIR. The next two panels map the APPRIS annotations onto the amino acid sequences of all coding variants and make the annotations visible in the genome regions provided by the UCSC Genome Browser (39). Finally, there are panels that allow the user to compare and contrast proteomic (37,40) and RNAseq (4) evidence tracks against the APPRIS annotation tracks in the UCSC Browser (see Supplementary Data for more details).

APPRIS has been designed to be portable, modular and flexible and it can be accessed as web services. These services can retrieve the results of the execution of APPRIS methods and other useful information for genes/transcripts. Plain text, JSON/GTF format or BED format (which facilitates the visualization of annotation tracks across genomic regions) outputs are available. In addition, APPRIS supports the downloading of data through the highly customizable BioMart (41) data mining tool.

APPRIS uses a MySQL relational database to store the information that can be downloaded from APPRIS web site. A comprehensive set of application programming interfaces (APIs) serve as a middle layer between underlying database schemes. The APIs encapsulate the database layout by providing efficient high-level access to data tables and isolate applications from data layout changes.

APPRIS ANNOTATIONS

APPRIS identifies a principal isoform for the majority of human genes. APPRIS determines a principal isoform for 17 731 (85.7%) of the 20 687 protein-coding genes in the GENCODE 7 (Ensembl 62) release (Supplementary Figure S4). A total of 53 307 variants were tagged as alternative by the methods in APPRIS and 22 799 transcripts were identified as principal isoforms (the discrepancy between genes and transcripts is because many transcripts are only alternatively spliced in the 3′- or 5′-regions and are regarded as identical by APPRIS because they have identical coding sequences).

Many of the isoforms recognized by APPRIS as alternative are likely to have substantial changes to their structure and function. A total of 37 550 alternative splice variants (70.4% of the variants tagged as alternative) would lose important functional or structural information relative to the principal isoform. The conservative estimates from the APPRIS methods show that 15 087 variants (28.3%) would lose important functional residues, 31 169 alternative gene products (58.5%) would have damaged or lost Pfam functional domains and 26 955 alternative isoforms (50.6%) would lose a substantial part of their 3D structure.

More than 8175 of the annotated transcripts would lose at least one *trans*-membrane helix and 543 would have lost a signal sequence. Almost 50 000 alternative splice variants (49 899, 94.5% of variants tagged as alternative) were less conserved across related species than the principal variant (from the results of CORSAIR, CExonic or INERTIA).

The CCDS project (42) is identifying consistently annotated, high-quality protein-coding variants for the human genome. CCDS variants are annotated only when there is agreement between the three main public annotation resources (GENCODE/Ensembl, NCBI and UCSC). Although the CCDS project can annotate any number of variants for a gene, many genes have a single CCDS variant, a variant agreed upon by all annotation resources. A single CCDS variant is the closest thing to an APPRIS principal variant, therefore, we should expect to see high agreement between the APPRIS constitutional isoforms and the CCDS variants.

For those genes that have multiple isoforms and a single CCDS variant, APPRIS is in agreement with the CCDS variant 93.5% of the time. What is more, this rises to over 96% for the core modules. This compares to an agreement of just 79.2% for the strategy of selecting the longest isoform (see Supplementary Table S1).

Two examples (of many) serve to illustrate the utility of APPRIS in the selection of principal isoforms. In the first example (gene *DNAJC5G*), APPRIS disagrees with the CCDS annotation by selecting an isoform that is 16 residues shorter than the pair of protein sequence identical isoforms chosen as the single CCDS variant, as the Ensembl reference sequence and as the SwissProt display sequence (Figure 2A). The variant selected by APPRIS (DNAJC5G-004) has a better score in Matador3D (it maps better to the known 3D structures in the PDB) and has a conserved Pfam domain. In contrast, the longer sequences would have broken Pfam domains and 3D structure (Figure 2B). The extra exon in the CCDS variant generates a 16-residue insertion that would be likely to disrupt a 3D structure (Figure 2C) and a conserved Pfam domain (Figure 2D).

The second example concerns the *TP63* gene. There are two well-studied isoforms for this gene, TA-alpha (43) and deltaN (44). They are generated from different translation start sites and generate different N-terminals (Figure 3A). However, rather than elect one of these two, APPRIS gives the best score to variant TP63-013, a 582-amino acid protein. Although this result might be surprising at first glance, it is perfectly logical.

TP63 is annotated with 15 coding variants in GENCODE 7, and all but 4 (TA-alpha, deltaN, P63delta and TP63-013) are rejected as potential principal isoforms by the SPADE (Pfam domains) and Matador3D (3D structure) modules in APPRIS. P63delta and TP63-013 are generated from TA-alpha and deltaN, respectively, by a known GYNGYN splicing event (45) that results in a swap of five amino acids 'GTKRP' for a single alanine in a non-critical region of the protein (Figure 3A). CORSAIR (vertebrate sequence database information) separates these four variants based on alignments with isoforms from other species (Figure 3B). It turns out that there is more ortholog evidence for deltaN and its GYNGYN variant TP63-013 than there is for TA-alpha and P63delta (the deltaN splice event is conserved as far back as chicken and *Danio*). There is also more CAGE data support for the deltaN/TP63-013 translation start site (A. Frankish, personal communication).

In addition, CORSAIR selects TP63-013 ahead of deltaN because *Danio* isoforms in the sequence databases have the single alanine instead of the GTKRP motif. In fact, the 3D structure of this region of the protein has also been solved for the isoform with the single alanine, adding weight to the APPRIS selection.

These examples neatly demonstrates the process behind APPRIS and reinforce the idea that it is possible to designate a principal isoform based on protein features and evolutionary antiquity, even where two isoforms (or more as in the case of *TP63*) have clearly defined functional roles.

A

APPRIS Search gene/transcript:

Home > Report View BioMart | Help & Docs | Contact

ENSEMBL Gene: ENSG00000163793 (DNAJC5G)
Location: chr2:27498289-27504367
Class: protein_coding
Status: KNOWN

Annotated isoforms

The box shows the coding variants annotated by ENSEMBL. Principal functional variants (when defined) are marked with a green tick.

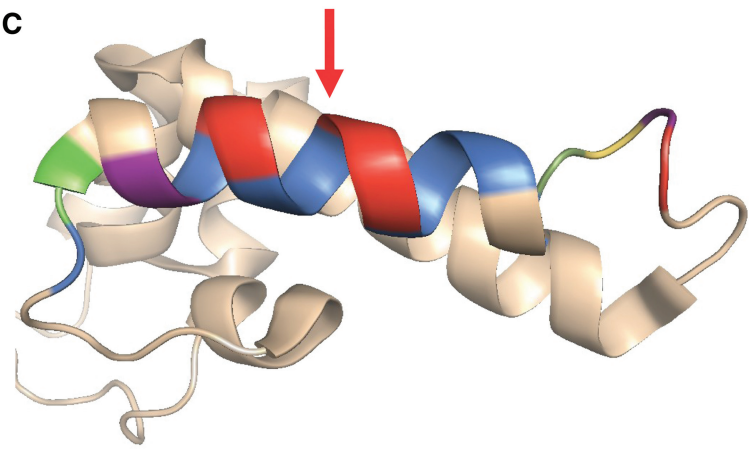
Show | Hide | Export

Transcript Id	Name	Class	Status	Length (bp)	Length (aa)	Codons not found	CCDS	Annotated Isoform
ENST00000296097	DNAJC5G-001	protein_coding	KNOWN	2008	189	-	CCDS1744.1	✗
ENST00000402462	DNAJC5G-002	protein_coding	KNOWN	1904	189	-	CCDS1744.1	✗
ENST00000404433	DNAJC5G-004	protein_coding	NOVEL	1647	173	-	-	✓*
ENST00000406962	DNAJC5G-003	protein_coding	NOVEL	1562	104	-	-	✗
ENST00000420191	DNAJC5G-007	protein_coding	NOVEL	593	62	stop	-	✗

B

Transcript Id	Status	Length (aa)	CCDS	Matador-3D	INERTIA	SPADE	Principal
DNAJC5G-001	**KNOWN**	189	**Yes**	0.75	**OK**	Damage	No
DNAJC5G-002	**KNOWN**	189	**Yes**	0.75	**OK**	Damage	No
DNAJC5G-004	NOVEL	173	-	**1.75**	**OK**	**Whole**	**Yes**
DNAJC5G-003	NOVEL	104	-	0.75	**OK**	Damage	No
DNAJC5G-007	NOVEL	62	-	0.75	**OK**	Damage	No

C



D

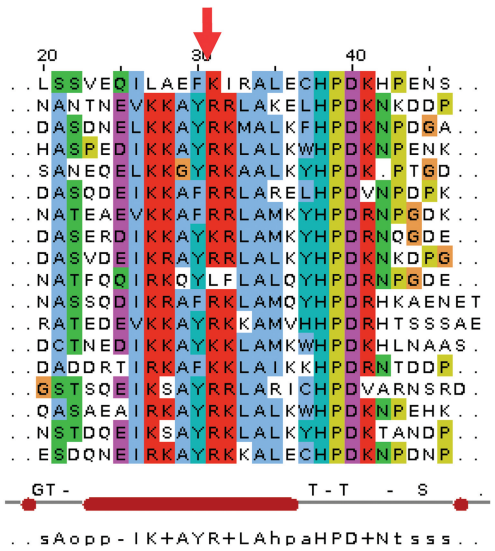
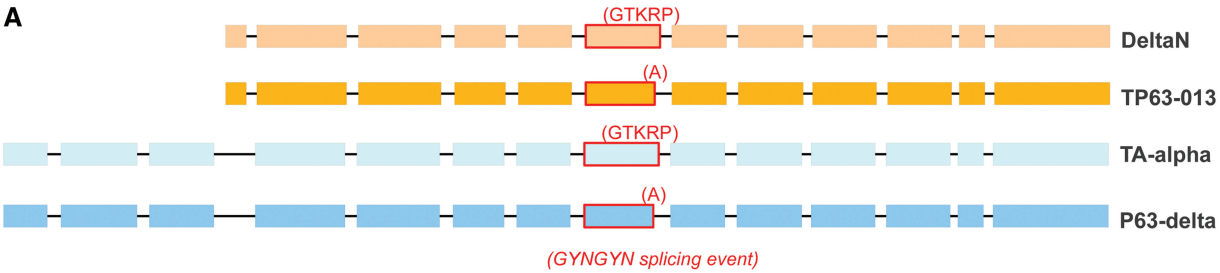


Figure 2. APPRIS annotations for gene *DNAJC5G*. (**A**) Snapshot of the APPRIS page for gene *DNAJC5G*, showing the five protein-coding transcripts annotated by Ensembl and the selection of the principal isoform by APPRIS (shown by the green tick). (**B**) A selection of the annotation results from the individual modules in APPRIS (Matador3D maps 3D structure to the isoforms, SPADE maps Pfam functional domains and INERTIA detects unusually evolving exons). The principal isoform is highlighted. APPRIS chooses isoform DNAJC5G-004 as the main variant based on the output of SPADE and Matador3D and designates the two 'KNOWN' isoforms (which are also CCDS variants) as alternative variants. (**C**) The 3D structure of mouse DNAJ subfamily C2 member 5 (PDB: 2CTW), to which DNAJC5G-004 has 56% identity with no gaps. The coloring on the 3D structure comes from the predominant coloring in the Pfam multiple alignment in (**D**). The large red arrow shows that the 16 extra residues present in the larger isoforms would have to be inserted into an important helix. (D) The multiple alignment for a section of the Pfam DnaJ family of sequences. The red arrow shows that the 16 extra residues in the CCDS variants would need to be inserted into a critical region of the functional domain of *DNAJC5G*.

A



B

Transcript	Alias	Length (aa)	CCDS	Matador-3D	SPADE	CORSAIR	Principal
TP63-004	**DeltaN**	586	**Yes**	**8.815**	**Whole**	5.5	No
TP63-013	-	582	**Yes**	**8.815**	**Whole**	**9**	**Yes**
TP63-001	**TA-alpha**	680	**Yes**	**8.815**	**Whole**	5.5	No
TP63-008	**P63-Delta**	676	**Yes**	**8.815**	**Whole**	7	No
TP63-010	**TA-beta**	555	**Yes**	7.3125	Damage	1	No
TP63-002	**TA-gamma**	487	**Yes**	6.8125	Damage	1	No

Figure 3. APPRIS result for gene *TP63*. (**A**) The four variants of gene *TP63* that score highest in APPRIS, highlighting the 5′-splice junction differences between TA and P63delta, deltaN and TP63-013, and the GYNGYN splice event that differentiates TA and deltaN from P63delta and TP63-013. (**B**) Annotation results from some of the individual modules in APPRIS (Matador3D maps 3D structure to the isoforms, SPADE maps Pfam functional domains and CORSAIR detects orthologous isoforms in related species). The isoforms are color-coded as in (A) and we have added two other well-known *TP63* variants to the table. APPRIS chooses isoform TP63-013 as the main variant based on the output of the three modules.

DISCUSSION

APPRIS deploys a range of computational methods to annotate alternative isoforms with protein structural and functional information and to evaluate cross-species conservation. The database provides reliable functional annotations for the most recent version of the manual annotation of the human genome. The APPRIS annotations will allow genome annotation groups and individual researchers to track the effect of alternative splicing events on individual splice isoforms.

There are already a number of databases that can annotate alternative transcripts with some of these features (46–48). What sets APPRIS apart from all these databases is that APPRIS provides high-quality annotations that are being used in the annotation of the human genome and that these annotations are used to select a principal isoform for each gene. Principal isoforms are selected based on evolutionary evidence in the form of conserved functional and structural motifs and cross-species conservation. The success of APPRIS is due to the observation that most alternative isoforms lack regions of conserved structure or function, or have exons that are evolving at measurably different rates compared with their principal counterparts. The APPRIS database has been able to identify a principal isoform for the majority of human genes (85%).

APPRIS is the first database to include principal isoforms on a genome-wide scale. Previously all database and large-scale studies have had to resort to selecting the largest annotated isoform as the reference variant. We have shown that this conservative solution is not ideal (18). The lack of reliably identified principal isoforms in annotation databases is an omission that is only going to become more glaring with time as the numbers of annotated variants in the sequence databases grow.

At present, most computational methods (49) and databases (21) are based on the assumption that a single isoform represents each gene. The SwissProt database, for example, combines all variants of the same gene in a single entry. These entries include experimental data and predictions, which are widely referenced from a number of external sources. One of the sequences in each entry, almost always the longest, is designated as the display sequence and the remaining sequences are included as splice variants. External databases and methods that use SwissProt as their standard often ignore these alternative sequences. If databases are going to condense gene products from the same gene into a single entry for technical reasons, it is better that the sequence that 'represents' the gene is the APPRIS principal isoform.

APPRIS principal isoforms have a wide range of uses and are applicable in all fields of research. Determining a principal isoform is important for research groups studying individual genes, since it is vitally important for designing experimental work. Researchers need to be able to work with the isoform that is most likely to have major functional activity, and this is not always clear for all genes. The designation of a single variant as the principal isoform is a critical first step for any genome analysis, for example, studies of cancer mutations (50) would be able to use APPRIS data to determine whether the mutations are principal or alternative exons, and proteomics studies could use APPRIS data to decide whether a peptide would be generated from on alternative or principal

exon. Since automatic prediction methods rely on the quality of input data, starting from the principal isoform should allow groups to perform more reliable studies. Finally, the selection of a principal isoform also serves as a starting point for investigations into the functional potential of alternative isoforms. These are just a few examples; the potential for the use of the APPRIS data in research is huge.

APPRIS is currently being used to annotate protein-coding genes by annotators in the GENCODE consortium (14) and the CCDS project (42). Annotations based on APPRIS data are already percolating to many users through these databases. We hope that the APPRIS principal isoforms will become accepted as the standard reference sequence for each gene. We believe that the principal isoforms identified by APPRIS are a significant advance on the current practice of selecting the longest variant as the reference isoform and that they should be used in all automatic genome-wide protocols and large-scale analyses.

The APPRIS annotations and the list of principal isoforms are accessible to all and are available for download in a range of formats.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–4, Supplementary Methods, Supplementary Results and Supplementary References [51–57].

ACKNOWLEDGEMENTS

The authors would like to thank Angel Carro and Edu Leon for invaluable technical assistance.

FUNDING

The Spanish National Institute of Bioinformatics (www.inab.org), a project of the 'Instituto de Salud Carlos III'; the Spanish Ministry of Science and Innovation [BIO2007-666855]; the ENCODE Project [U54 HG0004555]; Blueprint [282510]. Funding for open access charge: Spanish Ministry of Science and Innovation [BIO2007-666855].

Conflict of interest statement. None declared.

REFERENCES

1. Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 901–905.
2. Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.
3. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
4. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
5. Nicholson,P. and Mühlemann,O. (2010) Cutting the nonsense: the degradation of PTC-containing mRNAs. *Biochem. Soc. Trans.*, **38**, 1615–1620.
6. Weischenfeldt,J., Waage,J., Tian,G., Zhao,J., Damgaard,I., Jakobsen,J.S., Kristiansen,K., Krogh,A., Wang,J. and Porse,B.T. (2012) Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. *Genome Biol.*, **13**, R35.
7. Vasudevan,S., Peltz,S.W. and Wilusz,C.J. (2002) Non-stop decay—a new mRNA surveillance pathway. *Bioessays*, **24**, 785–788.
8. Harigaya,Y. and Parker,R. (2010) No-go decay: a quality control mechanism for RNA in translation. *Wiley Interdiscip. Rev. RNA*, **1**, 132–141.
9. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.-J., Yeats,C., Olason,P.I., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
10. Smith,C.W. and Valcárcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
11. Mudge,J.M., Frankish,A., Fernandez-Banet,J., Alioto,T., Derrien,T., Howald,C., Reymond,A., Guigó,R., Hubbard,T. and Harrow,J. (2011) The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.*, **28**, 2949–2959.
12. Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
13. Frankish,A., Mudge,J.M., Thomas,M. and Harrow,J. (2012) The importance of identifying alternative splicing in vertebrate genome annotation. *Database*, **2012**, bas014.
14. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference annotation for the ENCODE Project. *Genome Res.*, **22**, 1775–1789.
15. ENCODE Project Consortium, Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
16. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
17. UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource. *Nucleic Acids Res.*, **40**, D71–D75.
18. Tress,M.L., Wesselink,J.-J., Frankish,A., López,G., Goldman,N., Löytynoja,A., Massinghamm,T., Pardi,F., Whelan,S., Harrow,J. *et al.* (2008) Determination and validation of principal gene products. *Bioinformatics*, **24**, 11–17.
19. Lopez,G., Maietta,P., Rodriguez,J.-M., Valencia,A. and Tress,M.L. (2011) *firestar*—advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.
20. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B. and Westbrook,J.D. (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, 392–401.
21. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
22. Massingham,T. and Goldman,N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–1762.
23. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
24. Lassmannm,T. and Sonnhammer,E.L. (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.

25. Löytynoja,A. and Goldman,N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA*, **102**, 10557–10562.
26. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.
27. Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.
28. Käll,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
29. Viklund,H. and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1917.
30. Slater,G. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
31. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
32. Tress,M.L., Bodenmiller,B., Aebersold,R. and Valencia,A. (2008) Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol.*, **9**, R162.
33. Castellana,N.E., Payne,S.H., Shen,Z., Stanke,M., Bafna,V. and Briggs,S.P. (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl Acad. Sci. USA*, **105**, 21034–21038.
34. Chang,K.-Y., Georgianna,D.R., Heber,S., Payne,G.A. and Muddiman,D.C. (2010) Detection of alternative splice variants at the proteome level in *Aspergillus flavus*. *J. Proteome Res.*, **9**, 1209–1217.
35. Severing,E., van Dijk,A. and van Ham,R. (2011) Assessing the contribution of alternative splicing to proteome diversity in *Arabidopsis thaliana* using proteomics data. *BMC Plant Biol.*, **11**, 82.
36. Brosch,M., Saunders,G.I., Frankish,A., Collins,M.O., Yu,L., Wright,J., Verstraten,R., Adams,D.J., Harrow,J., Choudhary,J.S. *et al.* (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res.*, **21**, 756–767.
37. Ezkurdia,I., Del Pozo,A., Frankish,A., Rodriguez,J.M., Harrow,J., Ashman,K., Valencia,A. and Tress,M.L. (2012) Comparative proteomics reveals a significant bias towards alternative protein isoforms with conserved structure and function. *Mol. Biol. Evol.*, **29**, 2265–2283.
38. Alekseyenko,A.V., Kim,N. and Lee,C.J. (2007) Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA*, **13**, 661–670.
39. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
40. Desiere,F., Deutsch,E.W., Nesvizhskii,A.I., Mallick,P., King,N.L., Eng,J.K., Aderem,A., Boyle,R., Brunner,E., Donohoe,S. *et al.* (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.*, **6**, R9.
41. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.
42. Harte,R.A., Farrell,C.M., Loveland,J.E., Suner,M.M., Wilming,L., Aken,B., Barrell,D., Frankish,A., Wallin,C., Searle,S. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
43. Rouleau,M., Medawar,A., Hamon,L., Shivtiel,S., Wolchinsky,Z., Zhou,H., De Rosa,L., Candi,E., de la Forest Divonne,S., Mikkola,M.L. *et al.* (2011) TAp63 is important for cardiac differentiation of embryonic stem cells and heart development. *Stem Cells*, **29**, 1612–1683.
44. Crum,C.P. and McKeon,F.D. (2010) p63 in epithelial survival, germ cell surveillance, and neoplasia. *Annu. Rev. Pathol.*, **5**, 349–371.
45. Sinha,R., Lenser,T., Jahn,N., Gausmann,U., Friedel,S., Szafranski,K., Huse,K., Rosenstiel,P., Hampe,J., Schuster,S. *et al.* (2010) TassDB2—a comprehensive database of subtle alternative splicing events. *BMC Bioinformatics*, **11**, 216.
46. Birzele,F., Küffner,R., Meier,F., Oefinger,F., Potthast,C. and Zimmer,R. (2008) ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.*, **36**, D63–D68.
47. Shionyu,M., Yamaguchi,A., Shinoda,K., Takahashi,K. and Go,M. (2009) AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res.*, **37**, D305–D309.
48. Martelli,P.L., D'Antonio,M., Bonizzoni,P., Castrignanò,T., D'Erchia,A.M., D'Onorio De Meo,P., Fariselli,P., Finelli,M., Licciulli,F., Mangiulli,M. *et al.* (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res.*, **39**, D80–D85.
49. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Hériché,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
50. Quesada,V., Conde,L., Villamor,N., Ordóñez,G.R., Jares,P., Bassaganyas,L., Ramsay,A.J., Beà,S., Pinyol,M., Martínez-Trillos,A. *et al.* (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 47–52.
51. López,G., Valencia,A. and Tress,M.L. (2007) *firestar*—prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, W573–W577.
52. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
53. Castelo,R., Reymond,A., Wyss,C., Câmara,F., Parra,G., Antonarakis,S.E., Guigó,R. and Eyras,E. (2005) Comparative gene finding in chicken indicates that we are closing in on the set of multi-exonic widely expressed human genes. *Nucleic Acids Res.*, **33**, 1935–1939.
54. Remmert,M., Biegert,A., Hauser,A. and Söding,J. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM–HMM alignment. *Nat. Methods*, **9**, 173–175.
55. López,G., Valencia,A. and Tress,M.L. (2007) FireDB—a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.*, **35**, D217–D223.
56. Tress,M.L., Graña,O. and Valencia,A. (2004) SQUARE—determining reliable regions in sequence alignments. *Bioinformatics*, **20**, 974–975.
57. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
58. Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N. and Aebersold,R. (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–D658.